UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

September 21, 2000
DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-14

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donald Malec DM
Principal Researcher
Statistical Research Division

Subject: Accuracy and Coverage Evaluation Survey: Synthetic Assumptions

The attached document describes a general proposal for a report that we will prepare, per your request, following completion of applicable Accuracy and Coverage Evaluation Survey (A.C.E.) operations. The completed report is intended to aid the Executive Steering Committee on A.C.E. Policy (ESCAP) in its recommendation regarding the release of the statistically corrected data or the data without statistical correction as the P.L. 94-171 data. This report, together with other reports, will assess the operations and results of both the initial Census and the A.C.E. Both sets of assessments will be available to the ESCAP to aid the Committee in reaching its recommendation regarding the use of the statistically corrected data.

This report focuses on strategies for assessing the bias in small area estimates due to use of the synthetic estimator.

It is important to note that the conduct of the operations may lead us to modify the attached format by including additional information. It is also likely that descriptions and definitions will be enhanced or the data items could undergo revision. Conversely, we may conclude, for a variety of reasons, that some of the information set forth in the attached prototype may not be available. The attached document sets forth our conclusions prior to completion of the A.C.E. about what information would properly inform the ESCAP on this subject, but is subject to modification.

Accuracy and Coverage Evaluation 2000: Assessment of Synthetic Assumptions

Donald Malec

1. INTRODUCTION

The synthetic assumption holds that census coverage is homogeneous within a particular post-stratum. For example, the synthetic assumption implies that census counts in St. Louis, Missouri in a given post-stratum have the same relative odds of containing missed counts to erroneously enumerated counts as the census counts in the same post-stratum but in Milwaukee, Wisconsin. The synthetic assumption within post-strata will permit the Census Bureau to draw conclusions from the A.C.E. sample about the population as a whole, to individuals living in geographic areas not defined by post-strata. The synthetic assumption is necessary to permit precise statistical correction to small geographic areas based on a sample. The error that is introduced when the synthetic assumption does not hold is called synthetic bias.

Assessments of the 1990 PES were concerned with the possibility that synthetic bias introduced error in the PES, especially for low levels of geography such as blocks. Synthetic bias is of greater concern for small areas than for larger geographic aggregations. It is acknowledged that synthetic bias will likely result in the population of some blocks being overestimated and the population of other blocks being underestimated; statistical correction is not expected to produce unqualified improvement in the smallest geographic areas, like blocks.

While the accuracy of the A.C.E's synthetic estimates depends on the degree of homogeneity within post-strata, it is important to understand that perfect homogeneity cannot exist within all post-strata. The Census Bureau's evaluation of synthetic error should focus on whether heterogeneity of coverage is so great as to prevent an improvement from using the A.C.E., not on whether the post-strata are absolutely homogeneous. Additionally, the A.C.E. was designed to improve homogeneity as compared with the 1990 PES. The A.C.E. design has enhanced post-strata, including variables for mail return rate and type of enumeration areas.

This paper will present alternative methods to document and measure synthetic bias in the A.C.E. and the effects, if any, these violations had on the overall accuracy of the A.C.E., both numeric and distributive. The two components of synthetic bias, bias in the synthetic estimator and correlation bias, will be estimated separately and totaled. Synthetic bias will be measured at the Congressional district and state levels to best inform the ESCAP decision.

It is important to note that there is no generally-accepted method to measure synthetic bias in post enumeration surveys. Also, while the final document will assess the 2000 results in light of the 1990 data, direct comparison between these evaluations and the 1990 evaluations is not entirely possible because different formulas were used to evaluate the 1990 data.

2. OVERVIEW OF 1990 EVALUATIONS

Evaluations of synthetic estimates, using surrogate variables to create artificial populations of population counts have been documented in Fay and Thompson (1993), Freedman and Wachter (1994) and Kim, et al. (1995). In particular, Freedman and Wachter document a number of analyses using artificial populations. They provide estimates of the within post-strata and between post-strata variability; demonstrating within post-strata variability. A loss function analysis on the surrogate variables is also provided by Freedman and Wachter. Although the loss function analysis (on shares) is favorable to the use of the synthetic estimates (based on a census adjustment), it is pointed out the assumptions about the representativeness of the artificial populations are tentative and give variable results. In addition, Freedman and Wachter also show that loss function analysis using the synthetic estimate as the target may overstate the advantage of adjustment. This latter shortcoming is corrected to an extent, using some simplifying assumptions, by Fay and Thompson (1993) who perform a loss function analysis that incorporates both the artificial loss function of the synthetic estimator with a loss function that measures the other sources of bias and error in the DSE. In that analysis, the results are mixed. Kim et al. analyze state effects using both artificial populations and PES data. They also report mixed results. Heingartner and Speed (1993) analyze PES counts, at the block level and find heterogeneity beyond post-strata.

3. OVERVIEW OF METHODOLOGY

This section describes the essence of estimating synthetic bias. There are two components of synthetic bias - synthetic population bias and synthetic correlation bias.

3.1 Estimating the Synthetic Population Bias

The basic methodology used to estimate this component of synthetic bias is Artificial Populations.

Take a census variable (such as allocation rate) that is thought to be related to coverage and distribute the post-strata level undercount to small areas denoted by j in proportion to their size of the selected variable. For example, let

a_{ij} = the number of allocations in post-stratum i, area j

n_{cij} = the census count in post-stratum i, area j

u_i = the undercount in post-stratum i calculated from the post-stratum level DSE

N_{ij} = the "true" artificial population count in post-stratum i, area j

Then

$$N_{ij} = n_{cij} + a_{ij}\frac{u_i}{a_{i.}} \text{ where } a_{i.} = \sum_j a_{ij}$$

Thus $N_{i.} = n_{ci.} + u_i$ = the DSE for post-stratum i. Thus regardless of which census variable is used to create the artificial population, the post-stratum level total for the artificial population will be equal to the DSE for the post-stratum.

The population synthetic estimator for area j is:

$$\hat{N}_{.j} = \sum_i n_{cij} CCF_i$$

where $CCF_i = \dfrac{DSE_i}{n_{ci.}}$ is the coverage correction factor computed from the DSE for post-stratum i.

If for all areas j,

$$\frac{a_{ij}}{n_{cij}} = \frac{a_{i.}}{n_{ci.}}$$

then it is easy to show that $N_{.j} = \sum_i N_{ij} = \sum_i (n_{cij} + a_{ij}\frac{u_i}{a_{i.}}) = \hat{N}_{.j}$

Thus if, for example, the allocation rate for post-stratum i is homogeneous over areas j then the synthetic estimator will equal the true count for each area j. If the allocation rate is not homogeneous across small areas then $\hat{N}_{.j}$ will have a synthetic population bias estimated by:

$$SPB_j = \hat{N}_{.j} - N_{.j} \qquad (1)$$

This bias is caused by heterogeneity in the artificial population variable. If this bias is small and the artificial population variable is a good proxy for undercount, then the synthetic population bias of $\hat{N}_{.j}$ will be small.

3.2 Estimating Synthetic Correlation Bias

In addition to synthetic population bias, $\hat{N}_{.j}$ will have correlation bias caused by residual heterogeneity in the post-strata i. This correlation bias will be estimated using Demographic Analysis as described in Bell(2000). For each post-stratum, it will be distributed to the small areas in proportion to their census counts. For each post-stratum, let $\hat{D}_i$ denote the estimated correlation bias. The synthetic correlation bias for small area j is estimated by:

$$SCB_j \cdot \Sigma_i \frac{n_{cij}}{n_{ci.}}\hat{D}_i \qquad (2)$$

3.3 Estimating Total Synthetic Bias

The total synthetic bias , B $_j$, for small area j is the sum of equations (1) and (2), specifically

$$B_j \cdot SPB_j \cdot SCB_j \cdot (\hat{N}_{.j} - N_{.j}) \cdot \Sigma_i \frac{n_{cij}}{n_{ci.}}\hat{D}_i$$

This is the synthetic bias for examining numeric accuracy.

A bias term for the synthetic estimator of a population share, can be made using:

$$\frac{N_j + SPB_j + SCB_j}{\sum_j (N_j + SPB_j + SCB_j)} - \frac{N_j}{\sum_j N_j}$$

where N_j are constructed from the artificial population.

As mentioned by Freedman and Wachter (1994), there are any number of ways to produce an artificial population. Other ways of scaling census variables could be tried, as can ways that will create undercounts in some areas and overcounts in other areas. The preceding computations can be done for a number of artificial population variables. It can also be done separately for state and for congressional districts.

Some of the advantages of the method is that it is relatively straightforward to implement, given that other estimates, such as estimates of post-strata bias, are available. It gives an idea of the synthetic bias; holding everything else fixed. Large biases will be a warning. Some of the disadvantages are that the artificial variables may not be distributed like true coverage. The bias must still be put into context with other errors in the loss function. Small biases do not indicate that the synthetic assumption is harmless.

4. REFERENCES

Bell, William R. (2000). "Correlation Bias Results". DSSD Census 2000 Procedures and Operations Memorandom, Series B.

Fay and Thompson (1993). "The 1990 Post Enumeration Survey Statistical Lessons, in hindsight". *Proceedings of the 1993 Annual Research Conference.* U.S. Bureau of the Census, 71-91.

Freedman D., and K. Wachter (1994). "Heterogeneity and Census Adjustment for the Intercensal Base". *Statistical Science*, 476-485.

Heingartner, N. and T.P. Speed (1993). "Assessing between-block heterogeneity within the post-strata of the 1990 Post Enumeration Survey". *Journal of the American Statistical Association*, 88, 1047-1057.

Kim, J.J., A. Zaslavsky and R. Blodgett (1995). "Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census". *Survey Methodology*, 21 ,1 ,pp.55-62.